Genomics 97 (2011) 193–199

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments

L.J. McIver [a], J.W. Fondon III [b], M.A. Skinner [c,d], H.R. Garner [a,*]

[a] Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
[b] Department of Biology of the University of Texas at Arlington, TX, USA
[c] Department of Surgery of the University of Texas Southwestern Medical Center, Dallas, TX, USA
[d] Department of Surgery of Children's Medical Center, Dallas, TX, USA

ARTICLE INFO

Article history:
Received 17 September 2010
Accepted 3 January 2011
Available online 9 January 2011

Keywords:
Microsatellites
Repetitive DNA
Genomics
1000 Genomes Project

ABSTRACT

We performed an analysis of global microsatellite variation on the two kindreds sequenced at high depth (~20×–60×) in the 1000 Genomes Project pilot studies because alterations in these highly mutable repetitive sequences have been linked with many phenotypes and disease risks. The standard alignment technique performs poorly in microsatellite regions as a consequence of low effective coverage (~1×–5×) resulting in 79% of the informative loci exhibiting non-Mendelian inheritance patterns. We used a more stringent approach in computing robust allelotypes resulting in 94.4% of the 1095 informative repeats conforming to traditional inheritance. The high-confidence allelotypes were analyzed to obtain an estimate of the minimum polymorphism rate as a function of motif length, motif sequence, and distribution within the genome.

Published by Elsevier Inc.

1. Introduction

Globally, scientists are developing new technology in hopes of sequencing a complete human genome for ~$1000. This is a drastic price reduction when compared to the first human assembly, completed in 2001 by the Human Genome Project, at a cost of approximately $1 billion [1]. Sequencing costs have been rapidly decreasing with this trend even surpassing Moore's Law, which describes the pattern of computational power doubling every 18 months. One company predicts that a human genome can be sequenced in 15 min for around $1000 in approximately three years [2]. Another company recently announced that the costs of sequencing a complete personalized genome at 30× will be reduced from $48,000 to $19,500 with an additional significant discount for those individuals with diagnosed diseases[3].

The decreased cost of second generation sequencing platforms has made possible a new venture, the 1000 Genomes Project, undertaken by an international consortium with the goal of completely sequencing at least 2000 human genomes. This large-scale sequencing project, estimated to cost anywhere from $30 to $50 million, recently announced it had completed its three preliminary pilot studies [4,5]. In one pilot study 180 samples were sequenced at a low depth (~3×) while in another the exons of approximately 1000 genes in 1,000 samples were sequenced at a very high depth [5]. In the remaining study a deep sequencing of two families was performed. These two kindreds of mother, father, and daughter are from Utah and Nigeria. The 1000 Genomes Project recently described these sequencing data, reporting that genetic variations in microsatellite regions, sequences of DNA typically defined as tandemly repeated copies of one to six base pair motifs, are difficult to accurately capture [6].

Microsatellites are highly variable, mostly as a result of replication slippage [7]. These alterations in polymorphic repeat sequences are associated with a large number of diseases and may occur in coding and non-coding regions of the genome. For example, hereditary nonpolyposis colorectal cancer (HNPCC) is a condition characterized by a high degree of microsatellite instability (MSI) and exhibits simple tandem repeat variations in multiple coding regions [8]. Other simple sequence repeat alterations are also linked with neurological disorders and contribute to normal variations in behavior [9]. In addition to cancer and neurological associations, microsatellites have been shown to affect phenotypic non-disease traits such as craniofacial variation in domestic dogs [10].

To date, more than 20 heritable diseases have been associated with specific variations in microsatellite loci [11]. For example, spinocerebellar ataxia type 8 (SCA8) is caused by expansion of a CTG repeat in an exon of *ATXN8OS* and prostate cancer risk is associated with a variable CAG repeat in the *androgen receptor* gene [12,13]. Despite their abundance in the genome and strong association with human disease, only a relative few microsatellites have been thoroughly studied. This lack of knowledge of repetitive sequence variation is due to the high cost of sequencing and the difficulty of capturing microsatellite regions en masse. Accurate methods to measure global microsatellite polymorphism are essential to uncover possible

* Corresponding author. Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Washington Street, MC0477, Blacksburg, VA 24061-0477, USA. Fax: +1 540 231 2606.
E-mail address: garner@vbi.vt.edu (H.R. Garner).

0888-7543/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.ygeno.2011.01.001

Table 1
Average sequencing depth for all genomic regions (high complexity and microsatellite regions) as function of second generation sequencing platform.

				Pilot 2 (full genome)			Pilot 3 (exons of ~1000 genes)		
Genome	Family	Relation	ABI SOLiD	Illumina	LS454 GS FLX	Total	Illumina	LS454 GS FLX	Total
NA12878	Utah	Daughter	4.8×	30.9×	9.3×	44.9×	335.0×	128.2×	463.2×
NA12891	Utah	Father	–	22.0×	–	22.0×	74.6×	159.1×	233.7×
NA12892	Utah	Mother	–	18.4×	–	18.4×	173.9×	55.3×	229.2×
NA19238	Nigeria	Mother	–	13.7×	–	13.7×	121.4×	381.6×	503.0×
NA19239	Nigeria	Father	–	16.7×	–	16.7×	462.1×	23.2×	485.3×
NA19240	Nigeria	Daughter	40.0×	21.2×	3.0×	64.2×	445.4×	346.2×	791.6×

Average global coverage by sequencing platform for the two kindreds of the 1000 Genomes Project pilot studies is shown. In the 1000 Genomes Project, pilot two completely sequenced the genomes at a high depth while pilot three targeted the exons of approximately 1000 genes which cover about1.4 million base pairs. The two kindreds were not sequenced in pilot one.

biomarkers especially since the known heritability of disease cannot be accounted for by single nucleotide polymorphisms (SNPs) alone.

2. Results

Our goal in this research project was to establish a robust, reliable set of microsatellite allelotypes from which we could begin to make observations regarding the underlying genetics and statistical distributions of microsatellite repetitive elements therein. We evaluated the two kindreds (mother, father, and daughter) from the 1000 Genomes Project pilot studies, as the known lineage enabled us to verify the quality of the alignments by confirming Mendelian inheritance at each informative locus. As we would expect, with the exception of a small fraction of spontaneous variants, most loci obeyed this traditional inheritance pattern. Further, these two families were sequenced globally at a high depth with different read lengths, allowing us to evaluate the effectiveness of different 'deep sequencing' strategies through microsatellite loci.

2.1. Coverage of microsatellite-spanning long reads exhibit significant variation

The majority of the sequence reads for the two kindreds were completed on an Illumina automated sequencing platform which produces reads up to 100 base pairs in length. Only the two daughters' genomes were sequenced on the ABI SOLiD sequencing device, generating reads of 35 to 50 base pairs long, and the LS454 GS FLX machine, with sequences exceeding 350 base pairs in length (Table 1). The genomes of these two families were sequenced in two of the three pilot studies of the 1000 Genomes Project. For pilot two, the two kindreds' genomes were sequenced at high global depths with resulting global coverage varying from 13.7× to 64.2×. Pilot three, which targeted the exons of approximately 1000 genes, also had high coverage in the six individuals with averages ranging from 229.2× to 791.6×. This depth was possible since all exons sequenced for this pilot study totaled only ~1.4 million base pairs in length, approximately 0.05% of the human genome. In this study, to effectively capture microsatellite variation we only used reads which were at least 45 base pairs. Long reads resulted in a total global sequencing depth ranging from 2.2× to 21.3× with 52.6× to 791.6× depth in those exons sequenced in pilot 3 (Table 2).

These high depths are drastically reduced when considering only the low complexity regions of the genome. For example, even though the Utah daughter's genome was sequenced globally at 21.1× depth, microsatellites in gene regions had a coverage of only 5.0× (Table 3). Overall, the global reduction in sequencing depth in all low complexity regions is approximately 2 to 5 times. When considering the sequencing of specific exons, coverage in targeted regions in the same individual drops by about half from 503.0× globally to 283.5× in microsatellite only regions. This reduction in coverage in microsatellite regions could be due to alignment software difficulties associated with reads of low complexity or caused by selective DNA shearing in the sequencing process; Microsatellites in general may be more susceptible to shearing, causing the majority of reads containing them to include little to no high complexity regions, since just recently it was reported that AT-rich repetitive sequences are known to be the majority of cancer chromosomal breakpoints [14].

To compute the effective microsatellite coverage, a total of 376,685 microsatellites of at least 12 base pairs in length with no more than 10% insertions, deletions, and mismatches were identified in the human reference genome, build 36, using Tandem Repeat Finder [15]. All monomer repeats and those in non-LTR retrotransposons (SINE, LINE, and ALUs) were filtered from this set (Supplementary Table 1). We found that between 48% and 96% of the 376,695 microsatellites were completely covered by at least a single read in individuals in the two kindreds, after applying BWA, SAMTools, and our own custom software to create alignments (Supplementary Table 2).

2.2. Consensus assemblies are unreliable for long repetitive sequences

Most microsatellites, with the exception of some disease associate tri-nucleotide repeats, follow traditional Mendelian inheritance [16]. This predictable pattern combined with their high rates of polymorphism makes them useful for genetic profiling and paternity testing. Since November 1997, the FBI has been using 13 core simple tandem repeats, STRs, to aid in criminal investigations [17]. A larger set (over 400) of highly polymorphic microsatellite markers, identified by the

Table 2
Average sequencing depth for all genomic regions (high complexity and microsatellite regions) as a function of average read length.

			Pilot 2 (full genome)			Pilot 3 (exons of ~1000 genes)		
Genome	Family	Relation	Short reads	Long reads (45 bp+)	Total	Short reads	Long reads (45 bp+)	Total
NA12878	Utah	Daughter	23.8×	21.1×	44.9×	410.6×	52.6×	463.2×
NA12891	Utah	Father	0.8×	21.3×	22.0×	–	233.7×	233.7×
NA12892	Utah	Mother	2.6×	15.9×	18.4×	0.8×	228.4×	229.2×
NA19238	Nigeria	Mother	11.5×	2.2×	13.7×	–	503.0×	503.0×
NA19239	Nigeria	Father	15.2×	2.6×	17.7×	2.7×	482.6×	485.3×
NA19240	Nigeria	Daughter	52.7×	11.5×	64.2×	–	791.6×	791.6×

Coverage is shown for the two kindreds for different read sizes with long reads as those which are at least 45 base pairs in length. Since pilot three was a target enrichment study, the coverage was calculated as the average coverage over the target regions. The pilot two coverage was calculated as the average effective coverage based on the size of the reads.

Table 3
Average sequencing coverage for microsatellites by genomic locality.

(A)

	Reference (build 36)	Utah father (NA12891)	Utah mother (NA12892)	Utah daughter (NA12878)
Region	Microsatellite count	Average coverage	Average coverage	Average coverage
Upstream	4148	3.1×	2.3×	4.6×
5′UTR	21,119	3.9×	2.6×	4.9×
Exon	3955 (185)	3.2× (143.2×)	4.1× (132.6×)	3.4× (31.9×)
Intron	126,377	4.4×	2.9×	5.3×
3′UTR	6699	4.6×	3.4×	6.2×
Downstream	2934	4.5×	3.7×	6.9×
In/near genes	165,232	4.3×	2.9×	5.3×
Intergenic	211,453	4.1×	2.6×	4.8×
Total	376,685	4.2×	2.7×	5.0×

(B)

	Reference (build 36)	Nigerian father (NA19239)	Nigerian mother (NA19238)	Nigerian daughter (NA19240)
Region	Microsatellite count	Average coverage	Average coverage	Average coverage
Upstream	4148	1.8×	1.2×	4.8×
5′UTR	21,119	1.0×	0.8×	4.5×
Exon	3955 (185)	2.5× (311.3×)	1.0× (283.5×)	6.5× (475.9×)
Intron	126,377	1.4×	1.1×	5.1×
3′UTR	6699	2.5×	1.6×	6.4×
Downstream	2934	2.9×	1.9×	7.3×
In/near genes	165,232	1.4×	1.1×	5.1×
Intergenic	211,453	0.8×	0.7×	4.3×
Total	376,685	1.1×	0.9×	4.6×

Tandem repeat finder was used to identify microsatellites of at least 12 bps with no more than 10% insertions, deletions, or mismatches. These microsatellites were further processed to remove any which were contained in retrotransposon repetitive elements (ALU, SINE, or LINE). Monomer microsatellites were also removed from this data set resulting in a total of 376,685 tandem repeats found in the human reference. The total coverage of reads, at least 45 base pairs or greater in length, which completely cover the 376,685 microsatellites, including their flanking sequences, are shown for both the Utah (A) and Nigerian (B) families based on their locations relative to genes. Upstream and downstream were defined as the 1000 base pair sequences flanking the transcription start and end points. The reads shown are from the 1000 Genomes Project pilot two (full genome sequenced) and pilot three (exons only sequenced) with the coverage for the 185 microsatellites included in the targeted regions of pilot three shown in parenthesis.

Marshfield Medical Research Foundation, is currently used in linkage analysis studies [18]. Both the FBI Core STR and the Marshfield markers are generally long, with Marshfield markers varying in length from 8 to 361 base pairs, with an average length of 138 base pairs. Thus, determining accurate allelotypes with current deep DNA sequencing technology is challenging for those platforms limited to short reads. Indeed, the consensus sequences created using SAMTools on the alignment files provided by the 1000 Genomes Project exhibited very little variation in the FBI Core STR and Marshfield markers for the two kindreds, with only ~12.4% differing from the human reference genome. In contrast, the Celera and Venter genomes showed 56.7% and 75.5% variation, respectively. Of the Marshfield Markers which varied from the reference sequence in at least one individual, approximately 79% did not conform to Mendelian inheritance. Because the 1000 Genome Project kindreds do not exhibit the expected variation in microsatellite sequences represented in the well-characterized FBI and Marshfield markers, we infer that the sequences of many microsatellites obtained through consensus SAM files are inaccurate.

2.3. A new microsatellite alignment method was created to reliably distinguish alleles

To address this, we developed software to calculate microsatellite lengths based on reads which completely covered the entire repetitive region plus flanking sequence. We obtained reliable genotypes for only 19 of the Marshfield markers due to their long lengths. Eleven of the 19 markers exhibited sequences that differed from the human reference sequence, and all but two followed traditional inheritance patterns.

Due to the low number of reliable genotypes found for the Marshfield markers, we decided to identify and analyze all informative high-confidence repeats which differ between the two parents, found in our new alignments for both the Utah and Nigerian families. High-confidence allelotypes were defined as microsatellites sequenced at more than 3× depth but no more than 3 times the average coverage per genome (30×) with at least two reads supporting each allele with no more than two alleles found in the new alignments. Using these new higher stringency rules, we found a total of 1095 informative microsatellites in the Utah family with 94.4% following traditional Mendelian inheritance. Due to lower effective coverage, the Nigerian family only had 85 informative microsatellites of which 97.7% followed traditional inheritance. Approximately 2.4% of those microsatellites not exhibiting Mendelian inheritance are presumably due to sequencing errors, including but not limited to allele non-amplification and base call errors [19,20], while others could follow non-traditional inheritance because they are true spontaneous variants. For example, replication errors arising from errors in the mismatch repair system, account for mutations with a frequency of 10^{-2} per locus per cell, with this rate increasing upwards of 1000× in a disease state where the mismatch repair system has been compromised [21]. The relative contributions of errors and spontaneous real genomic changes for microsatellites needs to be studied in more detail, for which only more extensive analysis of the full 1000 Genome Project data set and independent validation of a statistically relevant number of loci with first generation sequencing will enable the exact determination of the magnitude of each component, which is a subject for future study.

2.4. Distributions and characteristics of microsatellites within the genomes of two kindreds

The number of high-confidence allelotypes per genome varied proportionally with respect to the depth of sequencing with 75.0% of the 376,685 microsatellites passing the high-confidence allelotype test for the Utah daughter (at a 21.1× depth) and only 11.8% of all microsatellites passing for the Nigerian mother (at a 2.2× depth) (Supplementary Table 3). Considering only those microsatellites with

high-confidence allelotypes, approximately 1–2% variation was seen globally in the two kindreds when compared to the reference genome, build 36 (Table 4). As expected, microsatellites varied considerably less frequently in exons which are under elevated levels of selection pressure. All of the 7 variable microsatellites in exons which were homozygous, differed from the reference genome by three base pairs, e.g. were frame conservative, while in all other regions, variations of one or two base pairs were the most common, accounting for 54.3% of differences.

Of the 376,685 total microsatellites, 227,849 were pure tandem repeats, with zero insertions, deletions, or mismatches. In the Utah family, the family sequenced at the highest depth which also had the largest number of high-confidence allelotypes, approximately 1.2% of the pure microsatellites had variations in one or more individuals, whereas 1.5% of microsatellites that contained insertions, deletions, or mismatches differed from the reference genome. Considering only exons in the Utah family, there was an average of ~1% variation with 62.5% of these being pure repeats.

Globally, two of the four dimers (TG and TA) exhibited the most variation, accounting for 42.5% and 28.1% of all variability, respectively. TG is also the most common microsatellite globally contributing 19.9% of all 376,685 loci while TA is slightly less common (5.9%). CAG, the most common motif family in exons, represents ~20% of the 3599 microsatellites in exons and contributes to 41.7% of the variation seen therein.

A total of 16,602 RefSeq genes contained or were within 1000 base pairs of at least one microsatellite with an average of 22.7 microsatellites per gene. Microsatellites with high-confidence alleles varied in 2242 genes with an average of 1.4 microsatellites varying in each of these genes. For example, *Receptor-type tyrosine-protein phosphatase delta* (*PTPRD*), which contains 31 exons and is involved in the development of multiple human cancers, was associated with the most microsatellites (382) [22]. In total there were four tandem repeats in introns and seven in the 5′UTR which differed in at least one of the two kindreds. It is interesting to note that *PTPRD* has recently been shown to exhibit high microsatellite instability [23]. The *dystrophin* gene (*DMD*), which is the longest human gene, at over 2.4Mbps with 51 exons, contained a significant number of microsatellites which (6 out of a total 358) differed in at least one of the two kindreds [24]. Repetitive sequences in this gene are useful in prenatal diagnosis of Duchenne muscular dystrophy [25,26]. The three genes with the highest number of variable microsatellites were *CSMD1* (13 variable microsatellites), *PTPRD* (11 microsatellites), and *RBFOX1* (11 microsatellites). All of the 13 variable microsatellites in *CSMD1* are in introns while seven of the variable microsatellites in *RBFOX1* are in introns with the remaining in the 5′UTR. *CSMD1* is the third largest human gene containing 70 exons with loss of heterozygosity associated with multiple cancers and multiple homozygous deletions found in this gene in oral squamous cell carcinoma [27–29]; Mutations in *RBFOX1*, another one of the largest genes in the human genome, were found in individuals with mental retardation and epilepsy with copy number variations found in those affected with autism [30,31].

2.5. Strengths and limitations of this approach on the 1000 Genomes Project data

A total of approximately 252,000 robust allelotypes were characterized for the Utah daughter's genome, which was sequenced on all three second generation platforms at 21.1× depth. The Utah father's genome, sequenced at 21.3× depth on just the Illumina platform, allowed for characterization of approximately 285,000 microsatellites showing that repetitive sequences can be accurately captured by the Illumina platform as long as coverage is high enough and the sequencing reads are long enough. With the method proposed, we required reads to be at least 45 base pairs long to capture microsatellites at minimum 12 base pairs in length. Thus, sequencing of the Utah daughter's genome, generating usable reads of 45 to over 300 base pairs in length, produced allelotypes for microsatellites that were between 12 and 75 base pairs, with 75 base pairs the maximum microsatellite length in the original set identified in the human reference sequence. The Utah father's genome, sequenced only on the Illumina, generated reads at a maximum of 100 base pairs, resulted in allelotypes for microsatellites at a maximum of 63 base pairs in length.

Table 4
Computed microsatellite variation relative to the human reference genome.

(A)

		Utah father (NA12891)		Utah mother (NA12892)		Utah daughter (NA12878)	
Region	Reference (build 36)	Count	% Diff	Count	% Diff	Count	% Diff
Upstream	4148	2393	0.3%	2147	0.2%	2384	1.0%
5′UTR	21,119	15,303	0.4%	13,800	0.3%	13,762	1.1%
Exon	3955	2832	0.0%	2530	0.0%	2858	0.2%
Intron	126,377	97,375	0.4%	87,966	0.3%	85,547	1.1%
3′UTR	6699	5251	0.4%	4686	0.3%	4596	1.2%
Downstream	2934	2249	0.7%	1973	0.3%	1975	1.2%
In/near genes	165,232	125,403	0.4%	113,102	0.2%	111,122	1.0%
Intergenic	211,453	159,984	0.5%	143,340	0.3%	140,271	1.2%
Total	376,685	285,387	0.5%	256,442	0.3%	251,393	1.1%

(B)

		Nigerian father (NA19239)		Nigerian mother (NA19238)		Nigerian daughter (NA19240)	
Region	Reference (build 36)	Count	% Diff	Count	% Diff	Count	% Diff
Upstream	4148	726	0.3%	411	2.0%	2729	1.0%
5′UTR	21,119	4065	0.4%	2420	1.6%	15,447	1.1%
Exon	3955	1145	0.0%	672	0.0%	3106	0.1%
Intron	126,377	25,432	0.4%	15,967	1.6%	96,208	0.1%
3′UTR	6699	1527	0.2%	965	0.9%	5223	1.0%
Downstream	2934	733	1.0%	415	1.7%	2246	1.3%
In/near genes	165,232	33,628	0.4%	20,850	1.3%	124,959	0.8%
Intergenic	211,453	36,050	0.6%	23,503	2.2%	157,418	1.3%
Total	376,685	69,678	0.5%	44,353	1.8%	282,377	1.0%

The total number of microsatellites with high-confidence allelotypes, repeats sequenced at more than 2× and not more than 30× with a maximum of 2 alleles, are shown for the Utah (A) and Nigerian (B) families. Microsatellite variations in this table were computed using reads from two of the three 1000 Genomes Project pilot studies.

On average, for both genomes, over 80% of the microsatellites of length 12 to 29 base pairs had robust allelotypes (Table 5). The sequencing of the Utah father resulted in a higher number of allelotypes for shorter microsatellites and a higher number over all, with the Utah daughter displaying significantly more allelotypes of microsatellites in the range of 40 to 75 base pairs; this is caused by the read sequence differences of the platforms. As expected, due to the fact that longer pure microsatellites are more likely to be polymorphic, only 0.2% of the pure microsatellites of less than 20 base pairs in the Utah daughter were found to vary from the reference genome, whereas 1.1% of the microsatellites 20 base pairs and over were variable [32].

The majority of microsatellites allelotyped in this study were relatively small, under 30 base pairs, due in part to the limitations placed by the 1000 Genomes Project data set. In this data set most genomes were sequenced solely on the Illumina platform which until recently was limited to short reads of 35 to 75 bps. This read size limitation combined with the reduction of coverage seen in microsatellite regions greatly reduced the average depth of useable reads in low complexity regions. Sequencing conditions required for this method to work would be deep sequencing at a minimum of 15× as we have seen coverage decreases in microsatellite regions up to 5 fold, on a platform with long reads such as the Illumina or LS454 GS FLX.

Minimal coverage used with this method should slightly underestimate global microsatellite variation, as it is less likely to capture heterozygous alleles where one of the two varies from the reference sequence. For example, the Nigerian father showed 0.5% global variation, 21.0% of those which varied were heterozygous, and 75.7% of these variable microsatellites were sequenced at 3–4× depth. Low variation is also seen in the Utah father (0.5%) and mother (0.3%) with 12.2% and 26.3% of the microsatellites sequenced at 3–4× depth. In comparison the Utah daughter with 1.1% global variation has 89.4% of all robust allelotypes sequenced at greater than 5× and 47.3% of its total variation is due to heterozygous repetitive sequences. The outlier of this trend is the Nigerian mother with 1.8% global variation, only 24.8% of which are heterozygous, with 41.8% of the microsatellites sequenced at 3–4× depth. This could be due to the fact that the 1000 Genomes Project found many more SNPs for the Nigerian family (over 4.5 million) than the Utah family (approximately 3.6 million) [33]; thus the mother could possibly have a genome which is slightly more variable than the other genomes in this study, including the reference. Another possibility could be poor sequencing of this genome resulting in poor reads.

In general, as with this method, higher coverage is always preferable as it will result in higher quality alignments; though at this time, whole genome sequencing at a high depth is still a cost prohibitive approach to capture global microsatellite alleles for a large set of individuals. If coverage is not high enough or the reads are not long enough to completely sequence through the microsatellite region, this method will not produce any results for the affected reads. This is in contrast to current alignment tools, such as BWA, which will produce possibly inaccurate lengths for low complexity regions even if the coverage is high because the coverage is in effect not sufficient as it is due to mainly short reads which do not completely span the microsatellite.

3. Discussion

3.1. Microsatellite alignment success is dependent on sequencing platform

The sequencing platform, which influenced both read length and sequencing depth, was a significant factor in microsatellite alignment success. Those genomes sequenced at a high depth on the LS454 GS FLX (with its longer read length) had more alignments to large repetitive regions, those in excess of 60 base pairs, while the genomes sequenced on the Illumina, which contributed more reads for the entire project, resulted in more alignments for shorter microsatellites. Overall, microsatellite coverage, and thus alignment success was higher in exons. This is a consequence of targeted exon sequencing from the 1000 Genomes Project pilot three which increased the effective depth of 185 of the 3855 microsatellites found therein. Microsatellite length also contributed to alignment success rates, with shorter microsatellites more likely to be covered completely in a single read.

3.2. 1000 Genomes Project alignments are unreliable at most microsatellite loci

The reliability of the consensus sequences at microsatellite loci could be quantified by inspection of allelotypes, especially for established marker sets, because the individuals were related. A majority of the ~400 Marshfield markers had defined allelotypes in the consensus sequences which were created by running SAMtools on the alignment files provided by the 1000 Genomes Project. Of the microsatellites which varied from the human reference sequence, a significant portion (79%) did not follow traditional inheritance. Inheritance patterns were vastly improved when the new alignment procedure with rules designed to exclude low-confidence reads or coverage was employed, enabling us to create a set of high-confidence allelotypes, with, for example, 94.4% of the 1095 informative microsatellites in the Utah kindred following Mendelian inheritance.

The 1000 Genomes Project data set is assumed to have been intended to primarily find single nucleotide polymorphisms (SNP) and short indels (insertions and/or deletions) because a majority of reads are 75 base pairs or less. These are high quality data sets but are not useful for studying the majority of microsatellite loci in the genome because each microsatellite locus must be sequenced through flanking regions to obtain accurate allelotypes. SAM consensus

Table 5
Robust allelotypes by microsatellite length for genomes sequenced at ~21×.

		Utah father (NA12891)		Utah daughter (NA12878)	
Microsatellite length	Reference (build 36)	Count	% Diff	Count	% Diff
10–19	268,540	244,535	8.9%	206,562	23.1%
20–29	51,008	32,562	36.2%	28,882	43.5%
30–39	27,329	7308	73.3%	10,238	62.5%
40–49	19,385	920	95.3%	8893	77.0%
50–59	7267	60	99.2%	967	86.7%
60–69	2444	2	99.9%	267	89.1%
70+	712	0	100.0%	69	90.3%
Total	376,685	285,387	24.2%	251,393	33.3%

All microsatellites identified in the human reference sequence, totaling 376,685, are shown with the total number of corresponding robust allelotypes found in the two genomes sequenced at the highest depth, around 21×, with long reads (45 bp+). The percent difference indicates the amount of microsatellites found in the reference sequence which did not have robust allelotypes in the sequences of the Utah father and daughter.

created using the BAM files provided by the 1000 Genomes Project do not accurately capture microsatellite variation, because they do not take into consideration that reads which do not span both the repetitive and flanking regions are effectively irrelevant at those loci. This lack of use of the read termination, along with a majority of short reads in the 1000 Genomes Project, results in most microsatellites portrayed as the same length as the reference which provided the consensus sequence backbone. A larger percentage of long reads in the 1000 Genomes Project BAM files would partially alleviate this problem. However, our data indicate that the best solution is to use custom software to calculate microsatellite allelotypes based on read alignment positions, because even though a genome is sequenced at an extremely high depth, microsatellites completely spanned by a single read are not common.

3.3. Distributions and characteristics of microsatellites within the genomes of two kindreds

High-confidence allelotypes were found for only 11.8% to 75.8% of the microsatellite loci with a global rate of variation around 1% for all genomes. These robust allelotypes were found for mainly short microsatellites ranging from 12 to 30 base pairs as this method is limited by the length of the sequencing reads in that if only short reads are provided then only short microsatellites are able to be characterized. Thus, due to the majority of short reads in the 1000 Genomes Project data, we were only able to analyze short microsatellites. Therefore, this study conveys a low estimate of microsatellite variation and further studies should be completed on a much larger cohort with a larger set of microsatellites sequenced at high depth to confirm and refine these results. Since longer microsatellite loci that are of high purity (few bases that deviate from the repetitive motif) are more likely to be polymorphic, it is particularly important that more long reads be gathered, to obtain an accurate picture of global microsatellite variation at these positions [32].

4. Conclusion

Deep sequencing coverage, where average sequencing depth across the whole genome is high, was found to be much lower, by a factor of approximately 2 to 5, in microsatellite regions. This lack of coverage, coupled with software typically used to align high complexity regions, was found to be insufficient to guarantee reliable allelotypes in repetitive regions. After devising an appropriate set of custom quality control rules we were able to reliably determine alleles for over 250,000 microsatellites. For these alleles, we observed an overall global microsatellite variation of ~1% on average for the members of the two kindreds from Nigeria and Utah which were sequenced at a high depth (~20×–60×) for pilot two of the 1000 Genome Project. Thus, having established a reliable method of determining microsatellite variation which will enable extension as the data accumulates for all 1000 Genomes Project individuals.

5. Materials and methods

5.1. Identifying microsatellites in the human reference genome

The human reference sequence, NCBI Build 36.1, produced by the International Human Genome Sequencing Consortium, was used as a control when determining microsatellite variation [34]. Microsatellites were located in this genome using Tandem Repeat Finder allowing for repeats of at least 12 base pairs in length with at least 90% purity [15]. A total of approximately 1.2 million microsatellites were found. All monomers and tandem repeats in simple repeats (SINE, LINE, and ALU), were removed from the data set resulting in a total of 376,685 loci. Some of these microsatellites were associated with RefSeq genes, provided by the UCSC Genome Browser, using their genomic location [35]. The upstream and downstream regions were defined as 1000 base pairs from the transcription start and end points of each gene.

5.2. Identifying microsatellites sequences from the 1000 Genomes Project

The binary alignment map, BAM, files for each of 6 individuals from the two kindreds was downloaded from the 1000 Genomes Project site [33]. Using SAMtools, version 3.1, the BAM files were transformed into files of consensus sequences [36]. A custom Perl script created flat text files containing a single representative base pair for each position in the genome. These files were of the same format as the human reference (hg18) assembly. BLASTable databases were created from these files so that the 50 base pair flanking sequences of the microsatellites found in the reference sequence could be aligned to each of the genomes from the 1000 Genomes Project [37]. The exact alignment position for each microsatellite was the point at which the two flanking sequences were within 1000 base pairs, as this is the largest microsatellite we have identified in the human reference genome [37,38]. This was done to accurately pinpoint the starting and ending positions of microsatellites in the consensus sequence.

5.3. Identifying microsatellites in the Celera and Venter genomes

Human reference genome, Build 36.1, microsatellites were aligned to their corresponding microsatellites in the Celera and Venter assemblies by creating a BLASTable database for each genome [37]. The 50 base pair flanking sequences of each of the microsatellites found in the reference genome were BLASTed against the Celera and Venter databases to determine their corresponding locations.

5.4. Identifying robust allelotypes in the 1000 Genomes Project using current alignment tools and additional custom microsatellite alignment software

BAM files were constructed using BWA and SAMtools with only those raw 1000 Genomes Project reads which were at least 45 bps in length [36,39]. These raw reads had already passed the basic quality control checks run by the 1000 Genomes Project [6]. Custom software determined the reads aligned to each TRF reference sequence microsatellite in the BAM file. Next the repetitive region was located on that read to determine if there was enough flanking sequence on either side of this region to accurately measure the microsatellite length. Microsatellites were allowed to have at most 10% insertions, deletions, and mismatches. If the microsatellite was flanked by high complexity regions on either side, it was considered to be a possible length. Next all possible lengths obtained for a microsatellite were analyzed. If there were no more than two alleles found per microsatellite and each allelotype was supported by at least 2 reads but no more than 3 times the coverage of reads per allele, then this microsatellite was considered a robust allelotype.

5.5. Calculating average sequencing depth for each genome

Average sequencing depth for pilot two, full genome sequencing at high depth, was calculated with the following equation, $(R^{*}L)/G$, where R is the total number of reads, L is the length of the reads, and G is the size of the genome sequenced. Pilot three was completed through targeted sequencing, so the sequencing depth was calculated as the average depth of points in alignments created using BWA, in the target regions provided by the 1000 Genomes Project [6,33].

5.6. Identifying the FBI CORE STR and the Marshfield markers in the reference genome

The FBI has identified 13 Core Simple Tandem Repeats, STRs, used in forensic genomic analysis [17]. The 50 base pair flanking sequences for the STRs were aligned to the human reference genome using BLAST with the default settings [37]. The most recent set of Marshfield markers, Set 16, was aligned to the reference genome in the same manner using the e-PCR primers from NCBI [40]. A Perl script was written to calculate if microsatellites from a family set followed Mendelian inheritance.

Supplementary data to this article can be found online at doi:10.1016/j.ygeno.2011.01.001.

Acknowledgments

This work was funded by the Virginia Bioinformatics Institute director's funds and the National Institute of Health, National Human Genome Research Institute, 1000 Genomes Project Dataset Analysis Grant (T-55818-363-1). We would like to extend special thanks to Dominik Borkowski, David Bynum, Cristi Galindo and Renee Nester for their valuable technical assistance.

All authors have agreed to all the content in the manuscript, including the data as presented.

References

[1] J. Markoff, I.B.M. Joins Pursuit of the $1,000 Personal Genome, The New York Times, 2009.
[2] Biology 2.0, The Economist, A Special Report on the Human Genome, 2010, pp. 3–5.
[3] M. Herper, Your Genome is Coming, Journal, (2010).
[4] N. Siva, 1000 Genomes project, Nat. Biotechnol. 26 (2008) 256.
[5] Wellcome Trust Sanger Institute Press Release: 1,000 Genomes Project releases data from pilot projects on path to providing database for 2,500 human genomes, June 21, 2010, http://www.sanger.ac.uk/about/press/2010/100621.html.
[6] R.M. Durbin, G.R. Abecasis, D.L. Altshuler, A. Auton, L.D. Brooks, R.A. Gibbs, M.E. Hurles, G.A. McVean, A map of human genome variation from population-scale sequencing, Nature 467 (2010) 1061–1073.
[7] T.A. Brown, Genomes, BIOS Scientific Publisher, Ltd., Manchester, UK, 2002.
[8] E. Forgacs, J.D. Wren, C. Kamibayashi, M. Kondo, X.L. Xu, S. Markowitz, G.E. Tomlinson, C.Y. Muller, A.F. Gazdar, H.R. Garner, J.D. Minna, Searching for microsatellite mutations in coding regions in lung, breast, ovarian and colorectal cancers, Oncogene 20 (2001) 1005–1009.
[9] J.W. Fondon 3rd, E.A. Hammock, A.J. Hannan, D.G. King, Simple sequence repeats: genetic modulators of brain function and behavior, Trends Neurosci. 31 (2008) 328–334.
[10] J.W. Fondon 3rd, H.R. Garner, Detection of length-dependent effects of tandem repeat alleles by 3-D geometric decomposition of craniofacial variation, Dev. Genes Evol. 217 (2007) 79–85.
[11] J.R. Brouwer, R. Willemsen, B.A. Oostra, Microsatellite repeat instability and neurological disease, Bioessays 31 (2009) 71–83.
[12] L. Kadouri, Z. Kote-Jarai, D.F. Easton, A. Hubert, R. Hamoudi, B. Glaser, D. Abeliovich, T. Peretz, R.A. Eeles, Polyglutamine repeat length in the AIB1 gene modifies breast cancer susceptibility in BRCA1 carriers, Int. J. Cancer 108 (2004) 399–403.
[13] C.J. Cummings, H.Y. Zoghbi, Fourteen and counting: unraveling trinucleotide repeat diseases, Hum. Mol. Genet. 9 (2000) 909–916.
[14] G.S. Chandok, K.K. Kapoor, R.M. Brick, J.M. Sidorova, M.M. Krasilnikova, A distinct first replication cycle of DNA introduced in mammalian cells, Nucleic Acids Res. (2010).
[15] G. Benson, Tandem repeats finder: a program to analyze DNA sequences, Nucleic Acids Res. 27 (1999) 573–580.
[16] C.L. Valon, New developments in mutation reserach, Nova Science Publishers, New York, 2006.
[17] J.M. Butler, Genetics and genomics of core short tandem repeat loci used in human identity testing, J. Forensic Sci. 51 (2006) 253–265.
[18] L.J. Rasmussen-Torvik, J.S. Pankow, J.M. Peacock, I.B. Borecki, J.E. Hixson, M.Y. Tsai, E.K. Kabagambe, D.K. Arnett, Suggestion for linkage of chromosome 1p35.2 and 3q28 to plasma adiponectin concentrations in the GOLDN Study, BMC Med. Genet. 10 (2009) 39.
[19] D.J. Koorey, G.A. Bishop, G.W. McCaughan, Allele non-amplification: a source of confusion in linkage studies employing microsatellite polymorphisms, Hum. Mol. Genet. 2 (1993) 289–291.
[20] K.R. Ewen, M. Bahlo, S.A. Treloar, D.F. Levinson, B. Mowry, J.W. Barlow, S.J. Foote, Identification and analysis of error types in high-throughput genotyping, Am. J. Hum. Genet. 67 (2000) 727–736.
[21] A.J. Simpson, The natural somatic mutation frequency and human carcinogenesis, Adv. Cancer Res. 71 (1997) 209–240.
[22] S. Veeriah, C. Brennan, S. Meng, B. Singh, J.A. Fagin, D.B. Solit, P.B. Paty, D. Rohle, I. Vivanco, J. Chmielecki, W. Pao, M. Ladanyi, W.L. Gerald, L. Liau, T.C. Cloughesy, P.S. Mischel, C. Sander, B. Taylor, N. Schultz, J. Major, A. Heguy, F. Fang, I.K. Mellinghoff, T.A. Chan, The tyrosine phosphatase PTPRD is a tumor suppressor that is frequently inactivated and mutated in glioblastoma and other human cancers, Proc. Natl Acad. Sci. USA 106 (2009) 9435–9440.
[23] P. Mokarram, K. Kumar, H. Brim, F. Naghibalhossaini, M. Saberi-firoozi, M. Nouraie, R. Green, E. Lee, D.T. Smoot, H. Ashktorab, Distinct high-profile methylated genes in colorectal cancer, PLoS ONE 4 (2009) e7012.
[24] D. Duan, From the smallest virus to the biggest gene: marching towards gene therapy for duchenne muscular dystrophy, Discov. Med. 6 (2006) 103–108.
[25] C. Oudet, R. Heilig, A. Hanauer, J.L. Mandel, Nonradioactive assay for new microsatellite polymorphisms at the 5′ end of the dystrophin gene, and estimation of intragenic recombination, Am. J. Hum. Genet. 49 (1991) 311–319.
[26] M. Maheshwari, R. Vijaya, M. Kabra, S. Arora, S.S. Shastri, D. Deka, A. Kriplani, P.S. Menon, Prenatal diagnosis of Duchenne muscular dystrophy, Natl Med. J. India 13 (2000) 129–131.
[27] D.I. Smith, Y. Zhu, S. McAvoy, R. Kuhn, Common fragile sites, extremely large genes, neural development and cancer, Cancer Lett. 232 (2006) 48–57.
[28] C. Ma, K.M. Quesnelle, A. Sparano, S. Rao, M.S. Park, M.A. Cohen, Y. Wang, M. Samanta, M.S. Kumar, M.U. Aziz, T.L. Naylor, B.L. Weber, S.S. Fakharzadeh, G.S. Weinstein, A. Vachani, M.D. Feldman, M.S. Brose, Characterization CSMD1 in a large set of primary lung, head and neck, breast and skin cancer tissues, Cancer Biol. Ther. 8 (2009) 907–916.
[29] C. Toomes, A. Jackson, K. Maguire, J. Wood, S. Gollin, C. Ishwad, I. Paterson, S. Prime, K. Parkinson, S. Bell, G. Woods, A. Markham, R. Oliver, R. Woodward, P. Sloan, M. Dixon, A. Read, N. Thakker, The presence of multiple regions of homozygous deletion at the CSMD1 locus in oral squamous cell carcinoma question the role of CSMD1 in head and neck carcinogenesis, Genes Chromosom. Cancer 37 (2003) 132–140.
[30] J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y.H. Lee, J. Hicks, S.J. Spence, A.T. Lee, K. Puura, T. Lehtimaki, D. Ledbetter, P.K. Gregersen, J. Bregman, J.S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M.C. King, D. Skuse, D.H. Geschwind, T.C. Gilliam, K. Ye, M. Wigler, Strong association of de novo copy number mutations with autism, Science 316 (2007) 445–449.
[31] C.L. Martin, J.A. Duvall, Y. Ilkin, J.S. Simon, M.G. Arreaza, K. Wilkes, A. Alvarez-Retuerto, A. Whichello, C.M. Powell, K. Rao, E. Cook, D.H. Geschwind, Cytogenetic and molecular characterization of A2BP1/FOX1 as a candidate gene for autism, Am. J. Med. Genet. B Neuropsychiatr. Genet. 144B (2007) 869–876.
[32] H. Ellegren, Microsatellites: simple sequences with complex evolution, Nat. Rev. Genet. 5 (2004) 435–445.
[33] 1,000 Genomes Project, May 1, 2010, http://www.1000genomes.org/.
[34] NCBI Human Reference Genome, Build 36.1, March 1, 2009, ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.36.1/.
[35] B. Rhead, D. Karolchik, R.M. Kuhn, A.S. Hinrichs, A.S. Zweig, P.A. Fujita, M. Diekhans, K.E. Smith, K.R. Rosenbloom, B.J. Raney, A. Pohl, M. Pheasant, L.R. Meyer, K. Learned, F. Hsu, J. Hillman-Jackson, R.A. Harte, B. Giardine, T.R. Dreszer, H. Clawson, G.P. Barber, D. Haussler, W.J. Kent, The UCSC Genome Browser database: update 2010, Nucleic Acids Res. 38 (2010) D613–D619.
[36] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.
[37] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[38] C.L. Galindo, L.J. McIver, J.F. McCormick, M.A. Skinner, Y. Xie, R.A. Gelhausen, K. Ng, N.M. Kumar, H.R. Garner, Global microsatellite content distinguishes humans, primates, animals, and plants, Mol. Biol. Evol. 26 (2009) 2809–2819.
[39] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics 25 (2009) 1754–1760.
[40] G.D. Schuler, Sequence mapping by electronic PCR, Genome Res. 7 (1997) 541–550.